MSR: Making Self-supervised learning Robust to Aggressive Augmentations

Anonymous Author(s)
Affiliation
Address
email

Abstract

Most recent self-supervised learning methods learn visual representation by contrasting different augmented views of images. Compared with supervised learning, more aggressive augmentations have been introduced to further improve the diversity of training pairs. However, aggressive augmentations may distort images' structures leading to a severe semantic shift problem that augmented views of the same image may not share the same semantics, thus degrading the transfer performance. To address this problem, we propose a new SSL paradigm, which counteracts the impact of semantic shift by balancing the role of weak and aggressively augmented pairs. Specifically, semantically inconsistent pairs are of minority and we treat them as noisy pairs. Note that deep neural networks (DNNs) have a crucial memorization effect that DNNs tend to first memorize clean (majority) examples before overfitting to noisy (minority) examples. Therefore, we set a relatively large weight for aggressively augmented data pairs at the early learning stage. With the training going on, the model begins to overfit noisy pairs. Accordingly, we gradually reduce the weights of aggressively augmented pairs. In doing so, our method can better embrace the aggressive augmentations and neutralize the semantic shift problem. Experiments show that our model achieves 73.1% top-1 accuracy on ImageNet-1K with ResNet-50 for 200 epochs, which is a 2.5% improvement over BYOL. Moreover, experiments also demonstrate that the learned representations can transfer well for various downstream tasks.

1 Introduction

A golden law in the context of computer vision is utilizing tremendous annotated data to learn good visual representations [53, 25]. Unfortunately, collecting annotated data with accurate labels is generally laborious, expensive [49, 51], and even infeasible [21]. To this end, various approaches have been proposed to learn such representations from unlabeled visual data, usually by performing visual pretext tasks. Among them, self-supervised learning methods [6, 7, 45, 9] based on contrastive loss have recently shown great promise, achieving state-of-the-art performance.

Representative contrastive methods are generally trained by maximizing agreement between differently augmented views of the same image (positive pairs), and increasing the distance between augmented views from different images (negative pairs) [48, 6, 16]. Compared with supervised learning, these works highlight the role of data augmentation for SSL and design more aggressive augmentation operations, such as grayscale, color jitter, and Gaussian blur. Although these aggressive augmentations can help to further improve the model performance, they also bring a severe semantic shift problem for training images. As illustrated in Figure 1, the first row shows original images from ImageNet [24] and CIFAR-100 [23] datasets. And the second row presents the corresponding augmented views with the widely used composition of augmentations [7, 6]. We can see that the

Submitted to 36th Conference on Neural Information Processing Systems (NeurIPS 2022). Do not distribute.

(a) Noisy samples from ImageNet-1K [24] (b) Noisy samples from CIFAR-100 [23]

Figure 1: The first row of Figure 1 are the original images (ImageNet-1K images are resized to squares) and noisy samples from aggressive augmentation are in the second row. From the first three images of (a), we can observe that Color jitter operation makes the image too bright or too dark that covers the details of images; and in the fourth and fifth column of (a), Gray and Gaussian blur operation leads images to be hardly distinguished from background; and the same augmentation strategy from ImageNet-1K leads to more vague images for CIFAR-100 shown in (b).

augmented views can be hardly recognized as semantically consistent with their original versions. Pushing these images to have similar representations can adversely affect the model training. Some recent works [56, 22] have also recognized this problem and resort to using weak augmentation to avoid it. However, directly discarding aggressive augmentations may reduce the diversities of training examples, resulting in limited representation ability. Therefore, in this paper, we retain the aggressive augmentations and try to counteract the subsequent semantic shift problem.

Specifically, we consider the semantically inconsistent pairs as noisy positive pairs. Since they are mixed with other semantic consistent pairs, it is hard to directly filter them out during training. Fortunately, recent works [2, 15, 52] show that deep neural networks (DNNs) have a crucial memorization effect that DNNs tend to first memorize clean (majority/semantically consistent) examples before overfitting noisy (minority/semantically inconsistent) examples. Motivated by this, we propose to set a relatively large weight for the aggressively augmented data pairs at the beginning of training to fully exploit all the training examples. And as the training goes on, the model begins to overfit semantically inconsistent data. Therefore we gradually reduce the weight of aggressive augmented pairs to neutralize their impact. Compared with ReSSL [56], our method can embrace the diverse examples from aggressive augmentations. And, compared with MoCo [16], SimCLR [6], BYOL [14], and few others, our method can significantly reduce the semantic shift problem.

Experiment results on multiple benchmark datasets show that our method can outperform state-of-the-art methods in various settings with a large margin. For instance, with 200 epochs of pre-training, our method achieves 73.1% Top-1 accuracy on ImagetNet-1K [24] linear evaluation protocol, which is 2.5% higher than BYOL [14]. Experiments on MS COCO [28] also show that our pre-trained model can continually improve the performance for multiple downstream tasks.

2 Related works

Self-supervised learning (SSL) has attracted great attention to capture universal representations [19, 55, 37, 47, 5, 11, 56]. The core of SSL is designing agent tasks, which allow us to learn representations from large-scale unlabeled data via pseudo labels instead of using any human annotations. To this end, many proposals devise different solutions in constructing pseudo labels, including predicting the rotation of images [12], putting pieces of images together [35], or recovering color from grayscale images [54]. Particularly, Wu et al. [48] propose an instance-level classification, which regards images augmented from the same image as a positive pair and others as negative examples. SimCLR [6] improves performance by inserting the projection network and introducing aggressive augmentation. He et al [16, 7] store negative representations in a queue to reduce the memory requirement. BYOL [14] enhances the power of SSL by removing the dependence of negative examples, which also addresses the problem of false negative examples [9]. Despite these methods have proved their effectiveness based on the aggressive augmentation, they ignore the semantic shift problem from it.

Noisy samples in aggressive augmentation. Recent studies [46, 39, 33, 38] have discovered that aggressive augmentation may generate noisy samples in the positive pairs. To alleviate this issue, ContraCAM [33] proposes a two-step approach to reduce the issue of random cropping, which seeks

objects first and then crops images based on their locations. In addition, Gansbeke et al. [44] conduct experiments on scene-centric datasets (e.g., COCO) containing multiple objects in images and argue that SSL can overcome the issue of random cropping. Unlike prior works focusing on random cropping, our work allows more general types of augmentation and can be viewed as a complement of previous studies.

Learning with noisy labels. Memorization effect of deep neural networks has been widely studied in the field of learning with noisy labels [2, 52]. In this regard, many state-of-the-art methods select confident examples by using first fit examples in the early learning phase [29, 36, 20, 34, 26]. Specifically, Co-teaching [15] uses the small-loss strategy to choose confident examples and employs two networks to reduce the bias from examples with noisy labels. JointOptim [40] selects confident examples based on the early stopping and replaces the labels of low confident examples to predictions. DivideMix [26] further improves Co-teaching [15] by regarding low confident examples as unlabeled data and training with semi-supervised learning techniques [3]. Although our method also leverages the memorization effect, existing works in learning with noisy labels focus on addressing large noise rate issues, which means that noise can be separated from training data by using early stopping trick. However, the noise rate in aggressive augmentation is very low, so we propose an indirect approach that does not extract noise to counter the noise impacts.

3 Methodology

Our approach aims to minimize the negative impacts of noisy positive pairs from aggressive augmentations while taking advantage of aggressive augmentations. As such, we first revisit preliminaries on self-supervised learning. Then, we elaborate on the proposed learning algorithm that counterbalances the noise impacts by utilizes memorization effects of DNNs.

3.1 Preliminaries on self-supervised learning

Self-supervised learning methods based on contrastive learning generally requires learning an embedding space that can easily separate different examples. Assuming D be a distribution of an image set. An image x is uniformly drawn from D. Denote t and t' as two different instances from the same distribution of image augmentation T. v and v' are two augmented views of the image x with $v = t(x)$ and $v' = t'(x)$, which are regarded as a pair of positive examples. Then, v and v' will separately feed through a encoder f_θ and a projector g_θ to embed z_1 and z_2, which are required to be close to each other via a contrastive loss function, e.g., InfoNCE [43] can be expressed as,

$$\mathcal{L}_{NCE} = -\log \frac{\exp(z_1 \cdot z_2/\gamma)}{\exp(z_1 \cdot z_2/\gamma) + \sum_{n \in N} \exp(z_1 \cdot n/\gamma)}, \tag{1}$$

where γ is a temperature parameter, and N is a set of negative example vectors. The embeddings of positive and negative examples are l_2-normalized. To stabilize the training process, some state-of-the-art methods [16, 56, 22] employ an asymmetric framework, including an online and a target networks. For the online and target networks, they have the same network structure but different weights of encoder f_ξ, and projector g_ξ, whose parameters ξ are updated by the online parameters θ with the exponential moving average method.

Recently, BYOL [14] finds negative examples are not necessary and adds a predictor q_θ in the online network to avoid collapsed solutions, e.g., all images have the same vector. And, the loss function can be simplified to,

$$\mathcal{L}_{mse} = 2 - 2 * \frac{\langle z_1, z_2 \rangle}{\|z_1\|_2 * \|z_2\|_2}. \tag{2}$$

3.2 Description of MSR

As discussed in Introduction, noisy positive pairs from aggressive augmentations will lead to severe semantic shift problem, which damages the generalization on downstream tasks. However, most of sample pairs generated from aggressive augmentations are beneficial for the model performance and

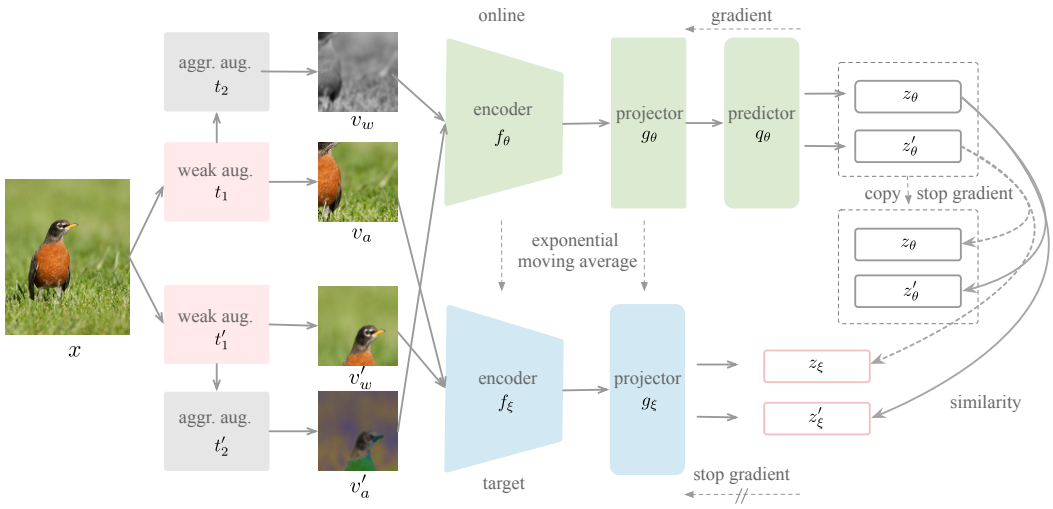

Figure 2: The illustration of our proposed method (MSR). We utilize an asymmetric-style framework, including an online and a target networks. The online network is optimized by gradients, and the target network is updated with the exponential moving average strategy. We first adopt the weak augmentation to generate two views (v_w, v'_w), then adopt the aggressive augmentations to further generate another two views (v_s, v'_s). Subsequently, we make aggressive-augmented views to keep consistent with their corresponding weak- and aggressive-augmented views in the embedding space. (Best viewed in color)

what precise degree of distortion would cause bias to remain unknown, so it is hard to distinguish from clean pairs and noisy pairs in SSL.

To this end, we propose an efficient method making self-supervised learning robust to the noise from aggressive augmentations, dubbed MSR. Envisioned by the memorization effects, DNNs will first fit the majority of training data in the early learning phase, and then overfit to the minority. Noisy pairs account for the minority, so we assume that the noise impacts will be small and then increases as the training process. We give different weights to clean and noisy pairs, and gradually reduce weights of noisy pairs to against raised noise impacts. Instead of separating noise from training data, we employ weak augmentation to generate sample pairs as relative clean pairs.

However, simply introducing clean pairs will significantly increase computation, and SSL has known that more computation will increases performance, which makes it difficult to fairly compare with other baselines. To limit computation, we meet two problems. First, generating more augmented instances, especially double instances, commonly means expensive computation, largely burdening CPU and hard disk, which leads to slow data processing down and low efficiency of GPU [6]. Second, more instances require more back-propagation times, which results in more processing time for GPU.

To address the first problem, we propose a novel data augmentation pipeline, called multi-stage augmentation, which can generate double instances with nearly the same resources. Specifically, data augmentation generates image variants by mixing different types of image transformation together and arranging into a queue. For an augmentation, it receives the image from the result of the previous augmentation, and transforms the image based on pre-defined probability, and then pass it to the next transformation, $t_{aug}(x) = t_2(t_1(x))$, where t_1 and t_2 are subset of t_{aug}. Based on this process, we can easily separate it into two child processes. Namely, t_1 includes weak augmentation operations while t_2 conducts aggressive augmentation operations based on the result of t_1. This process can be described as,

$$\begin{aligned} v_w &= t_1(x_i) \quad & v'_w &= t'_1(x_i), \\ v_a &= t_2(v_w) \quad & v'_a &= t'_2(v'_w). \end{aligned} \tag{3}$$

The noisy samples will have different impacts when occurring on different networks because of the asymmetric-style framework. When noisy samples occur on the online network and the target network receives clean samples, the target representation will help correct the online representation. If noisy samples occur on the target network, the online network will receive the wrong direction

Algorithm 1: MSR: Making Self-supervised learning Robust

Input: Neural network f_θ and f_ξ. Projector g_θ and g_ξ. Predictor q_θ. A batch of samples x. An image set D. Total number of training steps K. Weak augmentation function t_1. Aggressive augmentation function t_2. Hyper-parameter β_{base} in 5.

for $k = 1, \ldots, K$ **do**

  x is drawn from D

  $v_w = t_1(x) \quad v'_w = t'_1(x)$

  $v_a = t_2(v_w) \quad v'_a = t'_2(v'_w)$

  $z_\theta = q_\theta(g_\theta(f_\theta(v_a))) \quad z'_\xi = g_\xi(f_\xi(v'_w))$

  $z'_\theta = q_\theta(g_\theta(f_\theta(v'_a))) \quad z_\xi = g_\xi(f_\xi(v_w))$

  Update f_θ, g_θ and q_θ with loss function 2 and 5.

  Update f_ξ and g_ξ by slowly momentum update with the parameters of f_θ and g_θ

  Update β with equation 6

Output: The trained network f_θ

signal whether it receives clean samples or not. Therefore, we send aggressive augmented views into the online network and send weak augmented views into the target network.

$$\begin{aligned} z_\theta &= q_\theta(g_\theta(f_\theta(v_a))) \quad z'_\xi = g_\xi(f_\xi(v'_w)), \\ z'_\theta &= q_\theta(g_\theta(f_\theta(v'_a))) \quad z_\xi = g_\xi(f_\xi(v_w)). \end{aligned} \tag{4}$$

After obtaining four representations, we group them into four pairs, shown in Figure 2 and make each aggressive-augmented view to keep consistent with their corresponding weak- and aggressive-augmented views in the embedding space. For the second problem, instead of encoding aggressive augmented views on the target network, we use representations from the predictor due to the similarity of both. Accordingly, MSR only requires back-propagation twice, which is equal to many state-of-the-art SSL methods with a symmetrized loss [6, 8, 14, 5]. The total loss is summarized as,

$$\mathcal{L}_{total} = (1-\beta)\mathcal{L}_{mse}(z_\theta, z'_\xi) + \beta\mathcal{L}_{mse}(z_\theta, z'_\theta) + (1-\beta)\mathcal{L}_{mse}(z'_\theta, z_\xi) + \beta\mathcal{L}_{mse}(z'_\theta, z_\theta), \tag{5}$$

where β is a calculated parameter to re-weight the losses from weak and aggressive augmented samples. This loss function forces networks to learn representation that achieve a balance between weak- and aggressive-augmented views, which reduces overfitting to noisy pairs while avoiding a simple solution. To further against the raised noise impacts and increase the contribution of weak augmented samples, we decrease β with a cosine decay equation,

$$\beta = \beta_{base} \times \frac{1}{2}(cos(\pi\frac{k}{K}) + 1), \tag{6}$$

where β_{base} is a given number at the training beginning, and k and K are the current training steps and the total training steps.

4 Experiments

4.1 Datasets and Implementation details

Datasets: We evaluate our method on six image datasets, from small to large. We choose CIFAR-10/100 [23] for small datasets, and STL-10 [10] and Tiny ImageNet [1] for medium datasets, and ImageNet-100 [41] and ImageNet-1K [24] for large datasets. Note, ImageNet-100 contains 100 classes that are randomly selected from ImageNet-1K [24] and we choose the same classes with [41]. For STL-10, both 5k labeled and 100k unlabeled images are used for the pretrained model, and only 5k labeled images are used for the linear evaluation.

Table 1: Analysis of noise impacts of aggressive augmentations. We run methods on small and medium datasets for 200 epochs, and on ImageNet-100 for 100 epochs. The mean and standard deviation are computed over three trails.

Method	Aug.	β	CIFAR-10	CIFAR-100	STL-10	Tiny ImageNet	ImageNet-100
BYOL	AA	–	90.3±0.1	64.7±0.3	91.6±0.1	49.7±0.4	80.9
BYOL	AW	–	90.7±0.1	66.0±0.4	90.7±0.3	51.4±0.2	81.2
MSR	MA	Fixed	91.5±0.2	67.2±0.1	**93.0±0.1**	53.6±0.3	83.2
MSR	MA	Decay	**92.2±0.1**	**68.1±0.0**	**93.0±0.1**	**54.4±0.1**	**83.5**

Augmentation: In this paper, we define "aggressive" augmentation including random crop, horizontal flip, grayscale, color jitter and Gaussian blur, while "weak" augmentation includes random crop and horizontal flip. The hyper-parameters of the augmentations are following MoCo v2 [7] except the size of the cropped images for the small and medium datasets, and we resize images to 32×32 and 64×64 for the small and medium datasets, respectively.

Baselines: For the comparison, we re-implement the state-of-the-art methods, SimCLR [6], MoCo v2 [7], SimSiam [8] and BYOL [14] based on the public codes. We follow [8] that implements the MoCo v2 with a symmetrized loss function, and set the exponential moving average factor to 0.99 for all experiments. For BYOL [14] on the small and medium datasets, we follow [14], and set the channel inner layer of 1024 in the projection and prediction MLP and the output feature is 128. We set the exponential moving average factor beginning from 0.99 with a slowing increase to 1.

Network structure and optimization: Our method and reproduced methods are implemented by PyTorch v1.8 and we conduct all experiments on Nvidia V100. Our method is based on our reproduced BYOL [14]. The code of the proposed method is in the *supplementary material* and will be published after the paper is accepted.

For the pre-train stage on the small and medium datasets, we adopt ResNet-18 [18] as backbone. For optimization, we use SGD optimizer with a a cosine-annealed learning rate of 0.1 [30], and momentum 0.9, and a weight decay of 5×10^{-4} and a batch size of 256. We set $\beta_{base} = 0.3$ for CIFAR-10/100, and $\beta_{base} = 0.4$ for Tiny ImageNet and STL-10.

For the pre-train stage on large datasets, we adopt a standard ResNet-50 [18] as backbone. For ImageNet-100, the network is trained using SGD optimizer with a single cycle of cosine annealing [30], and initial learning rate of 0.2, and momentum 0.9, and a weight decay of 10^{-4}, and a batch size of 256. We conduct ImageNet-1K experiments with $8\times$ Nvidia V100 32G with Automatic Mixed Precision (AMP) package [31]. Specifically, we follow [14], and train a network with a LARS optimizer [50] with a single cycle of cosine annealing [30], and a momentum of 0.9, and a weight decay of 10^{-6}, and a batch size of 2048. The base learning rate starts from 0.9 and 0.6 for 100 and 200 epochs respectively, linearly scaled by the times of batch size 256 [13]. We set $\beta_{base} = 0.4$ for ImageNet-100 and ImageNet-1K.

Evaluation: We evaluate the representations of the pre-trained model with the linear evaluation protocol, which freezes the encoder parameters and trains a linear classifier on top of the pre-trained model. For the small and medium datasets, we follow the setting in MoCo v2 [7] and train a linear classifier for 100 epochs with an initial learning rate of 30, no weight decay, and a momentum of 0.9. The learning rate will be multiplied by 0.1 at the 60 and 80 epochs. For the large datasets, we follow the evaluation setting in Mean Shift [22], which only requires 40 epochs and a batch size of 256. The linear classifier is trained with SGD and an initial learning rate of 0.01, weight decay of 10^{-4} and a momentum of 0.9. The learning rate will be multiplied by 0.1 at 15, 30 and 40 epochs.

4.2 Preliminary Analysis

In this subsection, we investigate the influence of noise from aggressive augmentations on five datasets. We compare the linear performance of BYOL with two aggressive augmentations (AA), and with one aggressive and one weak augmentation (AW). Table 1 presents that BYOL with AW can largely improve the performance on CIFAR-100 and Tiny ImageNet, and mildly improve the performance on CIFAR-10 and ImageNet-100, but in STL, the network suffers from the weak augmentation. This inconsistent results verify our claim that aggressive augmentations have different degree of detrimental effects on different datasets. For those with more number of classes and small size images,

Table 2: Performance comparison with linear classification on small and medium datasets for 200 and 800 epochs. We adopt a ResNet-18 as backbone for all experiments. The mean and standard deviation are computed over three trails.

Method	CIFAR-10		CIFAR-100		STL-10		Tiny ImageNet	
	200 ep	800 ep	200 ep	800 ep	200 ep	800 ep	200 ep	800 ep
SimCLR	88.5±0.1	91.3	60.7±0.6	64.4	87.9±0.4	91.1	46.6±0.1	49.1
MoCo v2	86.9±0.2	90.8	60.5±0.3	65.0	88.3±0.4	91.2	47.6±0.2	50.7
SimSiam	87.6±0.2	91.6	56.2±1.2	62.3	85.7±0.3	89.7	41.0±0.3	43.9
BYOL	90.3±0.1	92.5	64.7±0.3	69.5	91.6±0.1	93.6	49.7±0.4	53.4
MSR (Ours)	**92.4±0.1**	**93.9**	**68.9±0.3**	**71.3**	**93.3±0.2**	**94.3**	**54.9±0.2**	**56.7**

Table 4: Performance comparison with linear classification on ImageNet-1K. All methods use a standard ResNet-50 as backbone without multi-crop strategy.

Method	Neg. pairs	Batch Size	Epochs	Top-1 Linear
Supervised		256	100	76.2
InstDis [48]	✓	256	200	56.5
PIRL [32]	✓	256	200	63.6
SimCLR [6]	✓	4096	1000	69.3
MoCo v2 [7]	✓	256	200	67.5
JCL [4]	✓	256	200	68.7
ReSSL [56]	✓	256	200	69.6
InfoMin Aug. [42]	✓	256	200	70.1
W-MSE 4 [11]		256	400	72.6
SimSiam [8]		256	200	70.0
SwAV [8]		4096	200	69.1
BYOL [8]		4096	100	66.5
BYOL [8]		4096	200	70.6
BYOL [14]		4096	300	72.6
MSR (Ours)		2048	100	71.4
MSR (Ours)		2048	200	**73.1**

it will have more negative impacts. By contrast, using weak augmentation may result in suboptimal results for datasets with large images and fewer classes.

Then, we compare BYOL with our proposed method MSR. In the third row in Table 1, we report the results of MSR with multi-stage augmentation (MA) and fixed β. MSR has improved linear classification across across five datasets, suggesting that aggressive augmentations can be leveraged while balancing negative effects. The results in the fourth row continue to improve by reducing β during the training process, which indicates that noise impacts vary at different stages of the training process and will become more apparent at the end. Notably, we set β_{base} as a default number of 0.5 in preliminary experiments, which means the weights of aggressive augmented instances and weak augmented are equal. Compared with Table 2 and 3, we can see the performance can be further improved with a turned β_{base}.

4.3 Linear Classification

Small and medium datasets. We first verify the effectiveness of our method on small and medium datasets, with short training time, 200 epochs and long training time, 800 epochs. We also run three trails for short running time to evaluate stability of the proposed method. Table 2 illustrates that the proposed method significantly improve the performance across the four datasets on short training time experiments, demonstrating that MSR can accelerate convergence and the small standard deviation also shows that the proposed method has good stability. For the long training time experiments, outstanding results show MSR can continue to improve the final performance.

Table 3: Performance comparison with linear classification on ImageNet-100 for 100 and 200 epochs. All methods adopt ResNet-50 as backbone.

Method	Batch size	100 ep	200 ep
SimCLR	256	79.1	82.4
MoCo v2	256	80.9	83.9
SimSiam	256	79.7	82.6
BYOL	256	80.9	83.6
MSR (Ours)	256	**83.7**	**85.5**

Table 5: **Transfer learning on downstream tasks:** object detection, instance segmentation and keypoint detection. All models pretrained on ImageNet-1K for 200 epochs and fine-tuned on MS COCO with 1 × schedule. Object detection and instance segmentation results are from [42] and [46] and keypoint detection results are from [46]. Results with * uses multi-crop strategy.

Method	Object detection			Instance segmentation			Keypoint detection		
	AP^{bb}	AP^{bb}_{50}	AP^{bb}_{75}	AP^{mk}	AP^{mk}_{50}	AP^{mk}_{75}	AP^{kp}	AP^{kp}_{50}	AP^{kp}_{75}
Random	32.8	50.9	35.3	29.9	47.9	32.0	63.5	85.3	69.3
Supervised	39.7	59.5	43.3	35.9	56.6	38.6	65.4	87.0	71.0
MoCo [16]	39.4	59.1	42.9	35.1	55.9	37.7	65.6	87.1	71.3
MoCo v2 [7]	40.1	59.8	44.1	35.3	55.9	37.9	66.0	87.2	71.4
InfoMin Aug. [42]	40.6	60.6	44.6	36.7	57.7	39.4	–.-	–.-	–.-
JCL [4]	–.-	–.-	–.-	35.6	56.2	38.3	66.2	87.2	**72.3**
SwAV* [5]	–.-	–.-	–.-	36.3	57.7	38.9	65.6	86.9	71.6
UOTA* [46]	–.-	–.-	–.-	36.7	58.4	39.4	**66.3**	**87.4**	**72.3**
MSR (Ours)	**41.1**	**61.4**	**45.1**	**37.3**	**58.6**	**40.1**	66.1	**87.4**	72.0

Table 6: **Transfer learning on downstream tasks:** object detection, and instance segmentation. All models pretrained on ImageNet-1K for 200 epochs and fine-tuned on MS COCO with 2 × schedule. Baseline results are from [42].

Method	Object detection			Instance segmentation		
	AP^{bb}	AP^{bb}_{50}	AP^{bb}_{75}	AP^{mk}	AP^{mk}_{50}	AP^{mk}_{75}
Random	38.4	57.5	42.0	34.7	54.8	37.2
Supervised	41.6	61.7	45.3	37.6	58.7	40.4
InstDis [48]	41.3	61.0	45.3	37.3	58.3	39.9
PIRL [32]	41.2	61.2	45.2	37.4	58.5	40.3
MoCo [16]	41.7	61.4	45.7	37.5	58.6	40.5
MoCo v2 [7]	41.7	61.6	45.6	37.6	58.7	40.5
InfoMin Aug. [42]	42.5	62.7	46.8	38.4	59.7	**41.4**
MSR (Ours)	**42.7**	**62.9**	**47.1**	**38.5**	**60.0**	**41.4**

Large datasets. We evaluate the performance of the proposed method on the large datasets, ImageNet-100 and ImageNet-1K. We reproduce all baselines on ImageNet-100 with batch size 256. The results on ImageNet-100 and ImageNet-1K are shown in Table 3 and Table 4 respectively. We can see that MSR outperforms the state-of-the-art methods on ImageNet-100 with a relatively large margin across 100 and 200 epochs. For results on ImageNet-1K, MSR consistently surpasses baselines, e.g., the performance of MSR for 100 epochs has already surpassed BYOL training for 200 epochs. MSR achieves a new state-of-the-art result for 200 epochs, exhibiting a 2.5% improvement over BYOL. Overall, empirical results on linear evaluation verify that MSR can accelerate the convergence rate and improve the generalization on various settings.

4.4 Transfer Learning

We further verify the quality of representation learned by MSR on more downstream tasks. For object detection and instance segmentation, we follow [42], and adopt Mask R-CNN [17] with FPN [27] to fine-tune our pretrained ResNet-50 model on COCO *train2017* with 1 × schedule and 2 × schedule, and evaluate performance on COCO *val2017*. Similar to object detection, we change Mask R-CNN to a keypoint version to conduct keypoint detection experiments with 1 × schedule. For more experimental details, please check the *supplementary material*.

Table 5 and Table 6 report the results of object detection, instance segmentation and keypoint detection. We can observe that MSR outperforms all baselines on object detection and instance segmentation tasks, especially, showing superior over the strong baseline InfoMin Aug. [42]. For the keypoint detection task, MSR also has comparable performance with UOTA, less than 0.3%. Note that UOTA employs multi-crop strategy with 8 views (2 × 224 + 6 × 96), which means although UOTA and MSR both train for 200 epochs, UOTA receives much more examples than MSR. Strong performance on downstream tasks demonstrates that MSR can improve the general quality of learned representation.

4.5 Training time comparison

We compare the running time between our method and reproduced BYOL. We conduct experiments on CIFAR-100 and STL-10 for 800 epochs with a single Nvidia V100, ImageNet-100 for 200 epochs with $4\times$ Nvidia V100, respectively. For ImageNet-100, we use Pytorch Automatic Mixed Precision package to speed up the training process and save GPU memory. Note that because we do not have the enough GPUs to run a standard BYOL with a batch size of 4096 on ImageNet, we use ImageNet-100 instead of ImageNet-1K, which has similar processing efficiency but ten times data.

Although our method uses double instances in the loss function 5, each instance in MSR only passes into the network one time. Therefore, the number of forward and backward keeps the same with BYOL. In addition, thanks to the efficient multi-stage augmentation, where we generate double instances using the nearly same resources. The main burden training time part of our proposed method may come from two more vector multiplication in our loss function 5. From Table 7, we can see that our method is as efficient as BYOL, and the differences between the two methods are less than 9% across three datasets.

Table 7: Training time comparison with BYOL on three datasets

Method	CIFAR-100	STL-10	ImageNet-100
BYOL	11.3h	77.6h	9.5h
MSR	11.5h	79.6h	10.4h

4.6 Ablation Studies

In this section, we investigate the sensitivity of the hyper-parameter β_{base} on three datasets. The linear classification results are shown in Figure 3. First of all, we can observe that MSR is not very sensitive to the value of β_{base}. MSR with $\beta_{base} = 0.3$ achieves the highest accuracy on CIFAR-100, and the best results on STL-10 and ImageNet-100 occur at $\beta_{base} = 0.4$. The performance with aggressive and weak augmented instances on the target network is higher than that with only weak augmented instances ($\beta_{base} = 0$), which suggests that only employing weak augmented samples on the target network may lead to suboptimal performance.

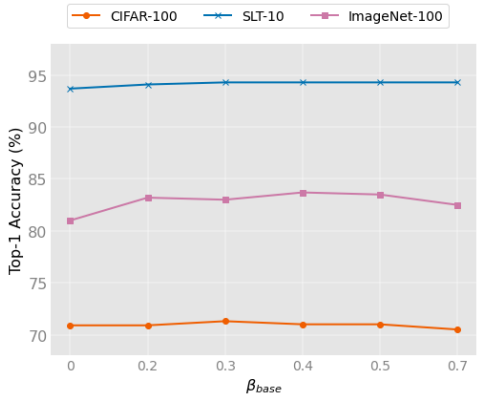

Figure 3: Sensitivity analysis for the hyper-parameter β_{base}. We conduct experiments on CIFAR-100 and STL-10 for 800 epochs, and ImageNet-100 for 100 epochs, respectively.

Furthermore, the value of β_{base} on medium and large datasets is higher than that on small datasets, which indicates that small datasets require higher weights to weak augmented samples to against noise impacts from aggressive augmented samples. This is consistent with our empirical finds that small datasets have higher probabilities of generating noise than medium and large datasets when employing the same strategy of aggressive augmentations.

5 Conclusion

In this paper, we first empirically demonstrate that two positive instances generated by the aggressive augmentations can cause the semantic shift issue, which introduces noisy positive pairs and degrades the quality of learned representation. To alleviate this issue, we propose a novel method Making Self-supervised learning Robust (MSR) to properly utilize aggressive augmentations and neutralize the semantic shift problem. Experimental results show that our proposed method achieves state-of-the-art results with the linear evaluation on various datasets and consistently improves the generalization on a series of downstream tasks.

The main limitation of this paper is that although we know there is noise from aggressive augmentations, we are still unknown the specific conditions under which noise occurs and how much networks suffer from it. We leave it as the future research direction.

References

[1] Zoheb Abai and Nishad Rajmalwar. Densenet models for tiny imagenet classification. *CoRR*, abs/1904.10429, 2019.

[2] Devansh Arpit, Stanislaw Jastrzebski, Nicolas Ballas, David Krueger, Emmanuel Bengio, Maxinder S. Kanwal, Tegan Maharaj, Asja Fischer, Aaron C. Courville, Yoshua Bengio, and Simon Lacoste-Julien. A closer look at memorization in deep networks. In *ICML*, pages 233–242, 2017.

[3] David Berthelot, Nicholas Carlini, Ian J. Goodfellow, Nicolas Papernot, Avital Oliver, and Colin Raffel. Mixmatch: A holistic approach to semi-supervised learning. In *NeurIPS*, pages 5050–5060, 2019.

[4] Qi Cai, Yu Wang, Yingwei Pan, Ting Yao, and Tao Mei. Joint contrastive learning with infinite possibilities. In *NeurIPS*, 2020.

[5] Mathilde Caron, Ishan Misra, Julien Mairal, Priya Goyal, Piotr Bojanowski, and Armand Joulin. Unsupervised learning of visual features by contrasting cluster assignments. pages 9912–9924, 2020.

[6] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey E. Hinton. A simple framework for contrastive learning of visual representations. In *ICML*, pages 1597–1607, 2020.

[7] Xinlei Chen, Haoqi Fan, Ross B. Girshick, and Kaiming He. Improved baselines with momentum contrastive learning. *CoRR*, abs/2003.04297, 2020.

[8] Xinlei Chen and Kaiming He. Exploring simple siamese representation learning. In *CVPR*, pages 15750–15758, 2021.

[9] Ching-Yao Chuang, Joshua Robinson, Yen-Chen Lin, Antonio Torralba, and Stefanie Jegelka. Debiased contrastive learning. *NeurIPS*, pages 8765–8775, 2020.

[10] Adam Coates, Andrew Y. Ng, and Honglak Lee. An analysis of single-layer networks in unsupervised feature learning. In *AISTATS*, pages 215–223, 2011.

[11] Aleksandr Ermolov, Aliaksandr Siarohin, Enver Sangineto, and Nicu Sebe. Whitening for self-supervised representation learning. In *ICML*, pages 3015–3024, 2021.

[12] Spyros Gidaris, Praveer Singh, and Nikos Komodakis. Unsupervised representation learning by predicting image rotations. In *ICLR*, 2018.

[13] Priya Goyal, Piotr Dollár, Ross B. Girshick, Pieter Noordhuis, Lukasz Wesolowski, Aapo Kyrola, Andrew Tulloch, Yangqing Jia, and Kaiming He. Accurate, large minibatch SGD: training imagenet in 1 hour. *CoRR*, abs/1706.02677, 2017.

[14] Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre H. Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Ávila Pires, Zhaohan Guo, Mohammad Gheshlaghi Azar, Bilal Piot, Koray Kavukcuoglu, Rémi Munos, and Michal Valko. Bootstrap your own latent - A new approach to self-supervised learning. In *NeurIPS*, pages 21271–21284, 2020.

[15] Bo Han, Quanming Yao, Xingrui Yu, Gang Niu, Miao Xu, Weihua Hu, Ivor Tsang, and Masashi Sugiyama. Co-teaching: Robust training of deep neural networks with extremely noisy labels. In *NeurIPS*, pages 8527–8537, 2018.

[16] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross B. Girshick. Momentum contrast for unsupervised visual representation learning. In *CVPR*, pages 9726–9735, 2020.

[17] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross B. Girshick. Mask R-CNN. In *ICCV*, pages 2980–2988, 2017.

[18] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, pages 770–778, 2016.

[19] Jyh-Jing Hwang, Stella X Yu, Jianbo Shi, Maxwell D Collins, Tien-Ju Yang, Xiao Zhang, and Liang-Chieh Chen. Segsort: Segmentation by discriminative sorting of segments. In *ICCV*, pages 7334–7344, 2019.

[20] Lu Jiang, Zhengyuan Zhou, Thomas Leung, Li-Jia Li, and Li Fei-Fei. Mentornet: Learning data-driven curriculum for very deep neural networks on corrupted labels. In *ICML*, pages 2309–2318, 2018.

[21] Davood Karimi, Haoran Dou, Simon K. Warfield, and Ali Gholipour. Deep learning with noisy labels: Exploring techniques and remedies in medical image analysis. *Medical Image Anal.*, 65:101759, 2020.

[22] Soroush Abbasi Koohpayegani, Ajinkya Tejankar, and Hamed Pirsiavash. Mean shift for self-supervised learning. In *ICCV*, pages 10306–10315, 2021.

[23] Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. Technical report, 2009.

[24] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *NeurIPS*, pages 1097–1105, 2012.

[25] Phuc H. Le-Khac, Graham Healy, and Alan F. Smeaton. Contrastive representation learning: A framework and review. *IEEE Access*, 8:193907–193934, 2020.

[26] Junnan Li, Richard Socher, and Steven C.H. Hoi. Dividemix: Learning with noisy labels as semi-supervised learning. In *ICLR*, 2020.

[27] Tsung-Yi Lin, Piotr Dollár, Ross B. Girshick, Kaiming He, Bharath Hariharan, and Serge J. Belongie. Feature pyramid networks for object detection. In *CVPR*, pages 936–944.

[28] Tsung-Yi Lin, Michael Maire, Serge J. Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft COCO: common objects in context. In *ECCV*, pages 740–755, 2014.

[29] Sheng Liu, Jonathan Niles-Weed, Narges Razavian, and Carlos Fernandez-Granda. Early-learning regularization prevents memorization of noisy labels. In *NeurIPS*, pages 20331–20342, 2020.

[30] Ilya Loshchilov and Frank Hutter. SGDR: stochastic gradient descent with warm restarts. In *ICLR*, 2017.

[31] Paulius Micikevicius, Sharan Narang, Jonah Alben, Gregory F. Diamos, Erich Elsen, David García, Boris Ginsburg, Michael Houston, Oleksii Kuchaiev, Ganesh Venkatesh, and Hao Wu. Mixed precision training. In *ICLR*, 2018.

[32] Ishan Misra and Laurens van der Maaten. Self-supervised learning of pretext-invariant representations. In *CVPR*, pages 6706–6716, 2020.

[33] Sangwoo Mo, Hyunwoo Kang, Kihyuk Sohn, Chun-Liang Li, and Jinwoo Shin. Object-aware contrastive learning for debiased scene representation. In *Advances in Neural Information Processing Systems*, pages 12251–12264, 2021.

[34] Duc Tam Nguyen, Chaithanya Kumar Mummadi, Thi-Phuong-Nhung Ngo, Thi Hoai Phuong Nguyen, Laura Beggel, and Thomas Brox. SELF: learning to filter noisy labels with self-ensembling. In *ICLR*, 2020.

[35] Mehdi Noroozi and Paolo Favaro. Unsupervised learning of visual representations by solving jigsaw puzzles. In *ECCV*, pages 69–84, 2016.

[36] Curtis G. Northcutt, Tailin Wu, and Isaac L. Chuang. Learning with confident examples: Rank pruning for robust classification with noisy labels. In *UAI*, 2017.

[37] Pedro O O Pinheiro, Amjad Almahairi, Ryan Benmalek, Florian Golemo, and Aaron C Courville. Unsupervised learning of dense visual representations. *NeurIPS*, pages 4489–4500, 2020.

[38] Xiangyu Peng, Kai Wang, Zheng Zhu, and Yang You. Crafting better contrastive views for siamese representation learning. In *CVPR*, 2022.

[39] Ramprasaath R Selvaraju, Karan Desai, Justin Johnson, and Nikhil Naik. Casting your model: Learning to localize improves self-supervised representations. In *CVPR*, pages 11058–11067, 2021.

[40] Daiki Tanaka, Daiki Ikami, Toshihiko Yamasaki, and Kiyoharu Aizawa. Joint optimization framework for learning with noisy labels. In *CVPR*, pages 5552–5560, 2018.

[41] Yonglong Tian, Dilip Krishnan, and Phillip Isola. Contrastive multiview coding. In *ECCV*, pages 776–794, 2020.

[42] Yonglong Tian, Chen Sun, Ben Poole, Dilip Krishnan, Cordelia Schmid, and Phillip Isola. What makes for good views for contrastive learning? In *NeurIPS 2020*, pages 6827–6839, 2020.

[43] Aäron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *CoRR*, abs/1807.03748, 2018.

[44] Wouter Van Gansbeke, Simon Vandenhende, Stamatios Georgoulis, and Luc V Gool. Revisiting contrastive methods for unsupervised learning of visual representations. 2021.

[45] Tongzhou Wang and Phillip Isola. Understanding contrastive representation learning through alignment and uniformity on the hypersphere. In *ICML*, pages 9929–9939, 2020.

[46] Yu Wang, Jingyang Lin, Jingjing Zou, Yingwei Pan, Ting Yao, and Tao Mei. Improving self-supervised learning with automated unsupervised outlier arbitration. 2021.

[47] Zhaoqing Wang, Qiang Li, Guoxin Zhang, Pengfei Wan, Wen Zheng, Nannan Wang, Mingming Gong, and Tongliang Liu. Exploring set similarity for dense self-supervised representation learning. In *CVPR*, 2022.

[48] Zhirong Wu, Yuanjun Xiong, Stella X. Yu, and Dahua Lin. Unsupervised feature learning via non-parametric instance discrimination. In *CVPR*, pages 3733–3742, 2018.

[49] Yan Yan, Rómer Rosales, Glenn Fung, Subramanian Ramanathan, and Jennifer G. Dy. Learning from multiple annotators with varying expertise. *Mach. Learn.*, 95(3):291–327, 2014.

[50] Yang You, Igor Gitman, and Boris Ginsburg. Scaling SGD batch size to 32k for imagenet training. *CoRR*, abs/1708.03888, 2017.

[51] Xiyu Yu, Tongliang Liu, Mingming Gong, and Dacheng Tao. Learning with biased complementary labels. In *ECCV*, pages 69–85, 2018.

[52] Chiyuan Zhang, Samy Bengio, Moritz Hardt, Benjamin Recht, and Oriol Vinyals. Understanding deep learning requires rethinking generalization. In *ICLR*, 2017.

[53] Daokun Zhang, Jie Yin, Xingquan Zhu, and Chengqi Zhang. Network representation learning: A survey. *IEEE Trans. Big Data*, 6(1):3–28, 2020.

[54] Richard Zhang, Phillip Isola, and Alexei A. Efros. Colorful image colorization. In *ECCV*, pages 649–666, 2016.

[55] Xiao Zhang and Michael Maire. Self-supervised visual representation learning from hierarchical grouping. In *NeurIPS*, pages 16579–16590, 2020.

[56] Mingkai Zheng, Shan You, Fei Wang, Chen Qian, Changshui Zhang, Xiaogang Wang, and Chang Xu. Ressl: Relational self-supervised learning with weak augmentation. 2021.

Checklist

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes]
   (b) Did you describe the limitations of your work? [Yes]
   (c) Did you discuss any potential negative societal impacts of your work? [N/A]
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes]
2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [N/A]
   (b) Did you include complete proofs of all theoretical results? [N/A]
3. If you ran experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [Yes]
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [Yes] See Section 4.1
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [Yes]
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [Yes] See Section 4.5
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
   (a) If your work uses existing assets, did you cite the creators? [Yes]
   (b) Did you mention the license of the assets? [N/A]

(c) Did you include any new assets either in the supplemental material or as a URL? [N/A]

(d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [N/A]

(e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [N/A]

5. If you used crowdsourcing or conducted research with human subjects...

(a) Did you include the full text of instructions given to participants and screenshots, if applicable? [N/A]

(b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A]

(c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A]